BlueGene/L System and Network Introduction

BlueGene/L Application Workshop

Katherine M Riley

July 27, 2005

Argonne National Lab

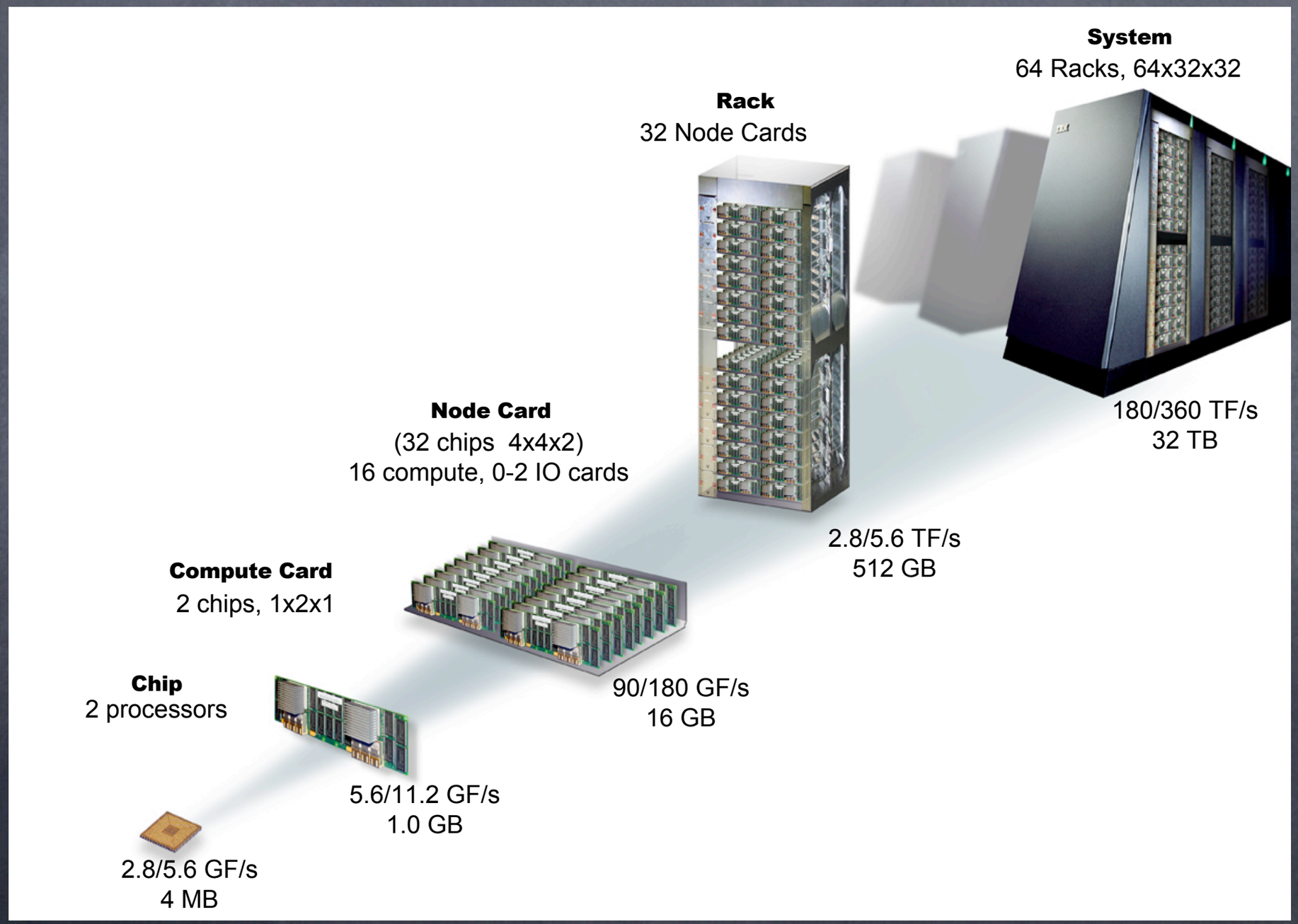
Intent

- A quick overview of the BlueGene/L system and networking with the target of aiding porting and optimizations

Idea Behind Design

- Massively parallel system
- Enough CPUS, must increase power efficiency
- Less individual CPU punch, cumulative a lot of punch
- More CPUs, so need a scalable network

Blue Gene System



CPU Information

- Two Processors per node
  - no SIMD between CPUs
    - Unsafe dual-proc programming model
    - Double FPU (double hummer)
- 32 bit architecture, 700 MHz

Memory

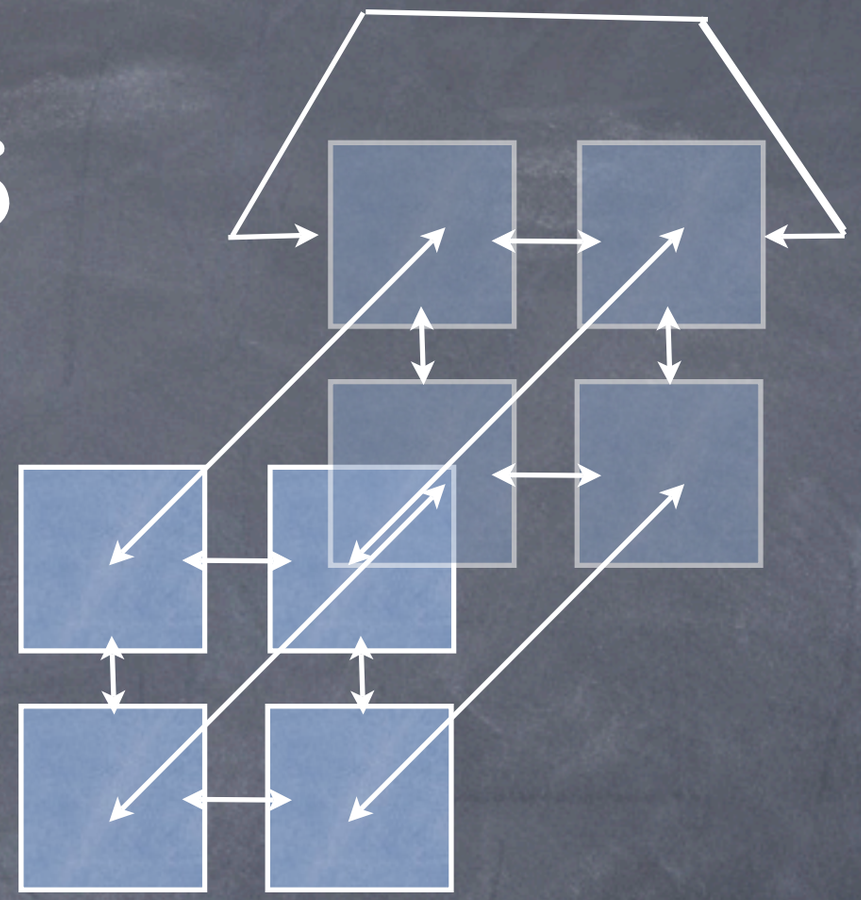
L1 Cache	32Kb, 32-byte line size
L2 Cache/Prefetch Buffer	16 128-byte lines
L3 Cache	4 MB ~35 cycle latency
Main Memory	512 MB DDR @ 350 MHz ~85 cycle latency

Compute Node Kernel

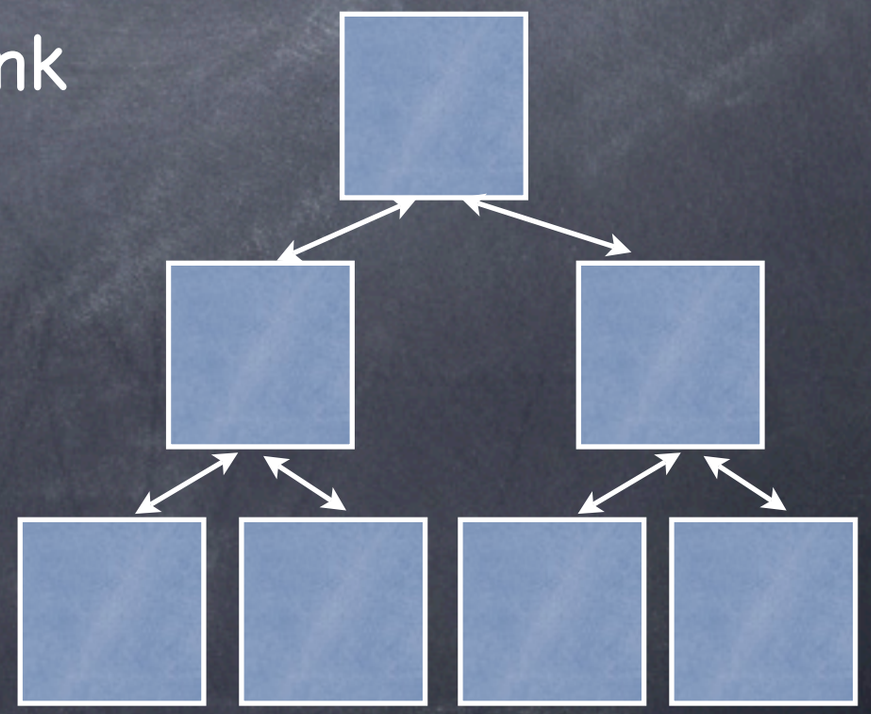
- Very small linux kernel on each CPU
  - 4 MB, can allocate 508MB
  - Less functionality
- Almost no noise
  - Excellent reproducable timings
  - Aids performance of global operations

Networks

- 3D Torus : point to point
  - Mesh with periodic boundaries
  - 1µs - 5µs latency, 1.4GB/s



- Global Tree : one to all
  - 2.5µs one way, 2.8GB/s per link
- Low Latency Barrier/Interrupt
  - 1.3µs roundtrip
- Ethernet and Control Network



Using the System

- Cross Compiling
  - Front end nodes compile for CNKs
    - No shell access to CNKs
- Executable loaded to PARTITION
  - Power of 2 in size
  - 512+ nodes for full torus

On the System

- IBM XL Compiler
- GNU C
- Developing collection of libraries
  - Math libraries
  - IO (pnetcdf, hdf5)

Running Jobs

- No dynamic libraries
- No threads
- Single executable per partition
  - Coprocessor or Virtual node mode
    - Coprocessor: 1 CPU for communication
    - Virtual Node mode: both to compute
  - Running:

```
mpirun -partition ANL_R001 -np 8 -cwd `pwd` -exe `pwd`/myJob
```

```
cqsub -t 60 -n 32 -m co `pwd`/myJob -env "BGLMPI_ALLREDUCE=MPICH"
```

Controlling the Network

- Can decide what network MPI routines use
  - -env 'BGLMPI_ALLREDUCE=TORUS'
- Dictate shape of partition
  - -shape NxNxN
- Dictate mapping of process
  - -env BGLMPI_MAPPING=TXYZ/XYZT
  - -mapfile filename

Things that are Different

- No dynamic libraries
- No threads
- Single executable
- Fixed Partition Size
- Generally smaller memory size